DOCUMENT RESUME

ED 322 190 TM 015 373

AUTHOR Stuck, Ivan A.
TITLE Parallel Merit Reliability: Error of Measurement from a Single Test Administration.
PUB DATE Apr 90
NOTE 18p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 17-19, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Error of Measurement; Estimation (Mathematics); *Research Methodology; Scoring; *Test Reliability
IDENTIFIERS Parallel Forms Reliability; *Parallel Merit Reliability; *Reliability Formulas

ABSTRACT

Parallel merit reliability (PMR) indexes the same consistency of measurement that is reflected in a validity coefficient; it reflects the reliability of measurement across identical merit score cases. Research has identified the potential benefits of the PMR approach as providing item level and cut-score reliability indices without requiring multiple administrations of the same test. Disadvantages of this type of reliability include: the need for a set of merit scores in addition to the item or interval scores; low magnitude coefficients; coefficients that are sensitive to the scale of the merit score; and the potential for artifactual bias when the merit intervals are too broad or of unequal discrimination. Computation of a parallel merit coefficient can be problematic if merit interval means are used as true score estimates. Means may be non-linear, single-case intervals may inflate the variance of means, and data processing may be awkward. However, by assuming linear relations between merit and item score, use of a regression approach may be justified, thereby, avoiding some problems associated with the use of interval means. A five-step process was used to investigate this approach to PMR. Two types of reliability coefficients were computed--parallel merit and coefficients representing coefficients of stability--for a real data set (high school graduation competency test data for 319 students in the 1988-89 school year) and a simulated data set (1,000 abilities generated to respond to 250 items ranging from the least to greatest ability level). Under the procedures presented, merit interval means were often non-linear and non-representative due to the number of interval cases. This preliminary effort to explore the usefulness of PMR lacked a credible set of real data. Future analysis should include a real data set having cases that match on pretest, posttest, and achievement test scores. Four data tables and three graphs are provided. (Author/RLC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

IVAN A. STUCK

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED322190

Parallel Merit Reliability:

Error of Measurement from a Single Test Administration

Ivan A. Stuck

Michigan State University

BEST COPY AVAILABLE

TM015373

Paper presented at the annual meeting of the National Council on Measurement in Education, 1990, Boston, Massachusetts

Parallel Merit Reliability:
Error of Measurement from a Single Test Administration

Ivan A. Stuck

Michigan State University

ABSTRACT

Parallel Merit reliability indexes the same consistency of measurement that is reflected in a validity coefficient; it reflects the reliability of measurement across identical merit score cases. An exploration of this under-acknowledged form of reliability (here called Parallel Merit reliability, or PM) has identified its potential benefits as providing item level and cut-score reliability indices without requiring multiple administrations of the same test. Disadvantages for this type of reliability include (1) the need for a set of merit scores in addition to the item or interval scores, (2) low magnitude coefficients, (3) coefficients which are sensitive to the scale of the merit score, and (4) the potential for artifactual bias when the merit intervals are too broad or of unequal discrimination. Computation of a PM coefficient can be problematic if using merit interval means as true-score estimates. Means may be non-linear, single-case intervals may inflate the variance of means, and data processing may be awkward. By assuming linear relations between merit and item score, however, use of a regression approach may be justified, thereby avoiding some problems associated with the use of interval means.

Parallel Merit Reliability:
Error of Measurement from a Single Test Administration

Ivan A. Stuck

Michigan State University

Reliability and error of measurement are both psychometric terms which refer to estimates of the consistency of instrument performance over some group of instrument administrations. Table 1 includes four of the five major types of reliability listed in Mehrens and Lehmann (1984), indicating how administrations are grouped for each reliability type. All types of reliability index consistency of measurement; nevertheless, types of reliability differ according to various patterns of administration. For coefficients of stability, consistency is observed across one or more administrations of the same test for each case. For coefficients of equivalence, consistency is observed across equivalent forms for each case. Consistency is indexed across sets of items in each case for internal consistency coefficients, and consistency is indexed across raters for each case for inter-rater reliability coefficients. For parallel merit (PM) reliability, however, consistency is indexed across cases having parallel merit.

A simple formula for reliability estimates is (1), the variance of true scores over the variance of observed scores (see

Table 1

Forms of Reliability

TYPE	GROUPING	ANALYSIS UNIT
Stability	Individual case	Individual
Equivalence	Individual case	Individual
Internal Consistency	Sets of items (scales)	Individual
Inter-rater	Individual work sample	Individual
Parallel Merit	Level of merit	Interval

Thorndike, 1988). In practical terms this translates to the following approximation:

$$\text{estimated reliability} = s^2_{means} / s^2_{scores} \qquad (1)$$

For most forms of reliability, save PM, the true scores (estimated by observed-score means) refer to the true scores of individual cases. For PM reliability, however, true scores are the expected values for parallel merit intervals (of cases falling into some particular score interval). Relative to conventional reliability coefficients, this can be expected to reduce the estimate of true score variance due to the absorption of interval points into the merit groupings; with administrations

nested within the individual (as in conventional reliabilities) there are potentially as many interval points as examinees.

The need for PM reliability stems from (1) the dearth of item level reliability indices, (2) a lack of alternative methods for estimating a standard error for an interval, and (3) as pointed out by Thorndike (1982), the value of differential reliability as an indicator of potential item bias. Although a coefficient of stability can be computed at the item level, often double test administrations are not feasible. A PM reliability procedure can provide an alternative means of computing an index of reliability for a single item.

Each type of reliability has its unique meaning; PM reliability's is the consistency of measurement within designated merit intervals. Its most popular form is the validity coefficient, a special case of PM reliability. Because it is a form of reliability, (in contrast to validity forms) it does not require a theoretical link between the merit levels and the instrument scores.

Its coefficients can be expected to vary according to the set of scores used to scale the merit intervals. One way to assure consistency is to regularly use a high quality, standardized instrument for scaling merit intervals. Otherwise, the test score totals may also serve as the basis for merit intervals.

Method

In this approach to PM reliability, a five step process was used: (1) score intervals were determined through the selection of an instrument to provide a merit score, (2) interval means were computed, (3) the variance of these means (associated with each case) was computed over all cases, (4) the variance of the scores was computed over all cases, and (5) the PM reliability was computed as the variance of means over the variance of scores. For a real and a simulated data set two types of reliability coefficients were computed: (1) PM, and (2) coefficients representing coefficients of stability.

Data. For an authentic data set, high school graduation competency test data (n= 319) was gathered for the 1988-89 school year, in the Grand Rapids (Michigan) School District for one of two administrations of a five item Life Skills subtest (essentially a test of forms completion). At least two raters were used to score the tests which happened to require a minimal amount of subjective judgment (e.g. was appropriate information supplied to the specified line?, etc). Two reliability coefficients were computed from the data: (1) an inter-rater coefficient (rater#1 scores were correlated with rater#2 scores to represent a coefficient of stability), and (2) a PM coefficient (rater#1 scores were used to compute a comparison set of PM coefficients). Raw scores from a recent standardized achievement test were used to order the merit intervals used in establishing the PM reliability.

A simulated data set was generated using a probabilistic IRT model (two parameter). N= 1000 abilities were generated to 'respond' to 250 items ranging systematically from the least to the greatest ability level (in difficulty). Ten of the mid-range items discriminated equally at a higher level than the remainder of the items (which discriminated equally). A 'retest' data set was similarly prepared by simply changing the seed values for the random functions used. By adding components of random error to each ability value, a set of simulated achievement scores was generated. This random error was created proportional to the standard error for a sixth-grade math achievement test.

Results

The real-data PM coefficients in Table 2 appear considerably smaller than their conventional counterpart (correlated ratings in column r_{rr}). This is not surprising, considering the fact

Table 2

PM reliability compared to the correlation of two rater scores

ITEM NUMBER	RELIABILITY PMr	PMm	PMrm	r_{rr}
1	.29	.33	.53	.91
2	.42	.37	.62	.85
3	.35	.30	.57	.77
4	.35	.41	.65	.95
5	.38	.41	.72	.91

PMr uses reading subtest for merit score
PMm uses math subtest for merit score
PMmr uses combined reading and math subtest for merit score

that the reading and math achievement subtest factors are remote from the life skills factors measured in the competency test. It is also likely that the concordance of raters may be somewhat inflated for this data set. By combining the reading and math achievement scores (to increase the number of intervals) the size of the coefficients is increased. Unfortunately, the number of single case intervals also increases, resulting in coefficient inflation. Because of the aforementioned problems with this data set, confidence intervals were not set about these coefficients.

For a set of simulated data, coefficients of stability appeared to have little relation to PM coefficients (r= .17 for coefficients in Table 3). The simulated PM coefficients,

Table 3

Simulated stability (S) and PM reliability

	ITEMS									
	I006	I012	I018	I024	I030	I036	I042	I048	I054	I060
S	.76	-.04	.32	.06	.50	.38	.52	.57	.72	.14
PM	.30	.15	.48	.18	.33	.29	.23	.25	.21	.19
	I066	I072	I078	I084	I090	I096	I102	I108	I114	I120
S	.18	-.05	.06	.24	.26	.33	.37	.27	.45	.40
PM	.20	.26	.34	.32	.36	.23	.46	.39	.53	.42
	I126	I132	I138	I144	I150	I156	I162	I168	I174	I180
S	.61	.62	.48	.17	.32	.17	.13	-.17	.46	.61
PM	.54	.37	.47	.30	.49	.33	.28	.26	.25	.32
	I186	I192	I198	I204	I210	I216	I222	I228	I234	I240
S	.23	.22	.34	.39	.72	.51	.84	.34	.85	.06
PM	.36	.25	.43	.34	.37	.38	.39	.34	.33	.37

r(S X PM)= .17

nevertheless, appear to be less variable than the simulated coefficients of stability.

In Table 4, confidence bounds (.95) are placed around the PM coefficients. Given an N of 1000 these coefficients appear reasonably precise. For evaluating potential item bias via differential reliability, a narrower confidence bound could be justified (p > .05).

Table 4

.95 confidence bounds on simulated PM coefficients

	ITEMS									
	I006	I012	I018	I024	I030	I036	I042	I048	I054	I060
UB	.386	.198	.618	.229	.427	.369	.295	.323	.270	.246
PM	.301	.154	.482	.178	.333	.288	.230	.252	.210	.192
LB	.235	.120	.376	.139	.260	.225	.179	.197	.164	.150
	I066	I072	I078	I084	I090	I096	I102	I108	I114	I120
UB	.251	.331	.436	.411	.461	.290	.586	.499	.676	.543
PM	.196	.258	.340	.321	.359	.226	.457	.390	.527	.423
LB	.152	.202	.265	.250	.280	.176	.357	.304	.411	.330
	I126	I132	I138	I144	I150	I156	I162	I168	I174	I180
UB	.690	.476	.604	.386	.621	.422	.354	.328	.317	.409
PM	.539	.372	.471	.301	.485	.329	.276	.256	.247	.319
LB	.420	.290	.367	.234	.378	.257	.215	.199	.193	.249
	I186	I192	I198	I204	I210	I216	I222	I228	I234	I240
UB	.463	.324	.545	.436	.477	.491	.500	.437	.429	.478
PM	.361	.253	.425	.340	.372	.383	.390	.341	.334	.373
LB	.282	.197	.331	.265	.290	.299	.304	.266	.261	.291

UB= upper bound
LB= lower bound

Discussion

Under the procedures presented, merit interval means were often non-linear and non-representative due to the number of interval cases. If the theoretically desirable assumption is allowed that true item scores and true parallel merit are linear, the single case interval problem and the non-linearity problem can be overcome by expressing the reliability as a regression coefficient (where merit value is the predictor and merit interval mean [of item score] is the criterion); single or low case-count intervals to be excluded from the parameterisation). If a reliability coefficient is desired, (2) can be used:

$$\text{estimated reliability} = s^2_{\text{estimates}} / s^2_{\text{scores}} \qquad (2)$$

If standardized regression is used, and the items and merit scores measure common factors, the resulting coefficient is a validity coefficient, a special case of PM reliability.

By adopting a regression approach, the door is opened to the use of additional factors as predictors. For example, regression studies of item bias can be regarded as studies of differential PM reliability, (testing the significance of achievement level and of group membership when entered as predictors of the criterion (item score or test score).

The following conditions should optimize the precision of PM reliability indexes:

(1) a large number of cases

(2) small size (or large number) of merit intervals

(3) the equality of merit intervals

(4) the quality and number of factors used to account for score variance.

Discrepant Discrimination Artifact Hypothesis

As stated before, the use of an ordering scale to control for merit in a partial correlation, or a multiple regression test for item bias, can be regarded as a study of differential parallel merit reliability (although it is more commonly discussed as differential validity). A dichotomous group membership variable controlled-for on merit may be used to predict item score. Where group membership correlates significantly with item score, the potential for item bias is acknowledged. The following alternative hypothesis identifies a potential cause of the apparent bias as a problem of unequal merit intervals:

Hypothesis: Merit intervals are apt to differ in levels of discrimination, thus encouraging the appearance of artifactual item bias. Such artifact is predictable when (1) item and test discrimination differ and (2) subgroups in the test population have different distributions.

Consider the following examples. Given subgroup A with a mean of 300 and subgroup B with a mean of 500 on a 1000 point scale, both having field test item scores and achievement test (merit) scores. The achievement test discriminates optimally at about 500 on a 1000 point scale. Item #1 discriminates well at about 300 on a 1000 point scale. There is a reasonable likelihood that the item may falsely appear to be biased against group A. Another item, #25, discriminates well at around 700 on a 1000 point scale. There is a fair likelihood that this item may appear biased against group B. The unfortunate reality is that a merit interval for group A is not actually equivalent to the same merit interval for group B. This problem is exacerbated by (1) poor discrimination for this merit interval (low test discrimination at this achievement or ability level), (2) the severity of distributional differences between groups A and B, and (3) the discrimination level of the item being examined. Figure 1 depicts a merit interval as it falls on overlapping A and B distributions. By viewing the shapes of the distribution slices, it becomes obvious that the group A interval consists of a larger proportion of lower merit cases than does the corresponding distributional slice for group B. As merit discrimination is lowered, the interval is increased in size (the bounds of the actual true scores are expanded), thereby increasing the discrepancy. The level of discrimination for the item under examination, simply improves the item's ability to separate lower merit level cases from higher merit level cases

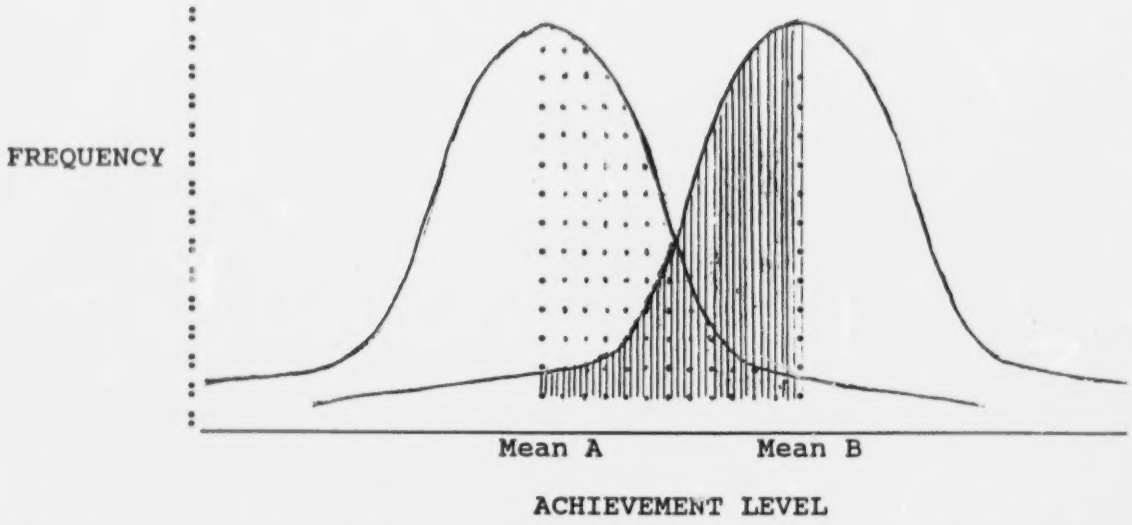

Figure 1. Two subgroups with differing distributions illustrate how their intervals may differ qualitatively despite being ordered on a similar achievement scale. Note the relative proportions of high performing examinees in the achievement interval between Mean A and Mean B.

which remain within the merit intervals despite the intentions of the psychometrician to control them. Figures 2 and 3 graph item means by merit level for two math items which have no apparent cultural content; item 137 appears biased against group A ($p < .001$), whereas item 144 appears biased against group B ($p < .001$). The discrimination of the achievement test being constant, the variation and discrepancy of item discrimination relative to the subgroup distributions may account for a substantial part of the apparent bias.

ITEM 137

Sam had $3.92, Bill had $.87, Ruth had $16.42, and Sue had $1.68.

How much did they have in all?

--> A. $22.89

B. $24.19

C. $109.02

D. $34.89

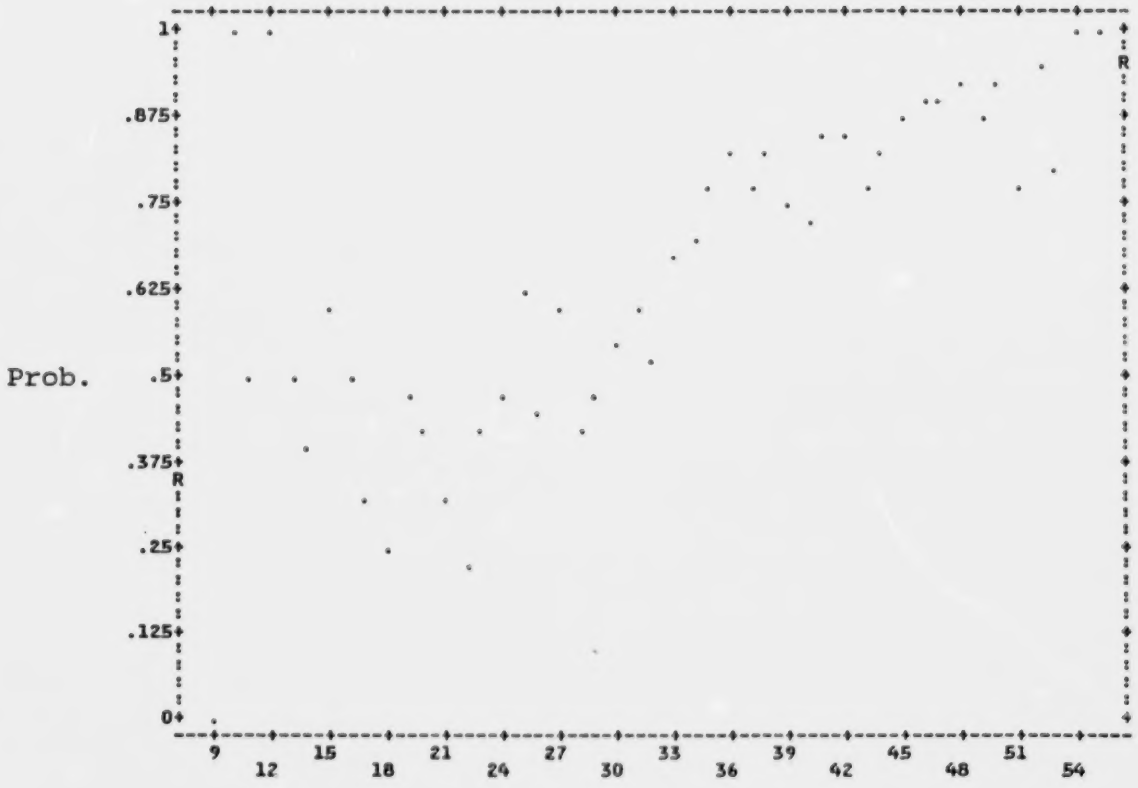

Figure 2. Proportion of correct response (p-value) by total score on a second math subtest indicates item 137 discrimination. Best discrimination is between math subtest scores 30 and 39. The item does not discriminate well at the lower end of the scale. Group A is significantly correlated ($p < .001$) with item score while controlled for achievement level.

ITEM 144

Which is the symbol for 3.14?

A) ≈

B) •||•

--> C) ♈

D) =

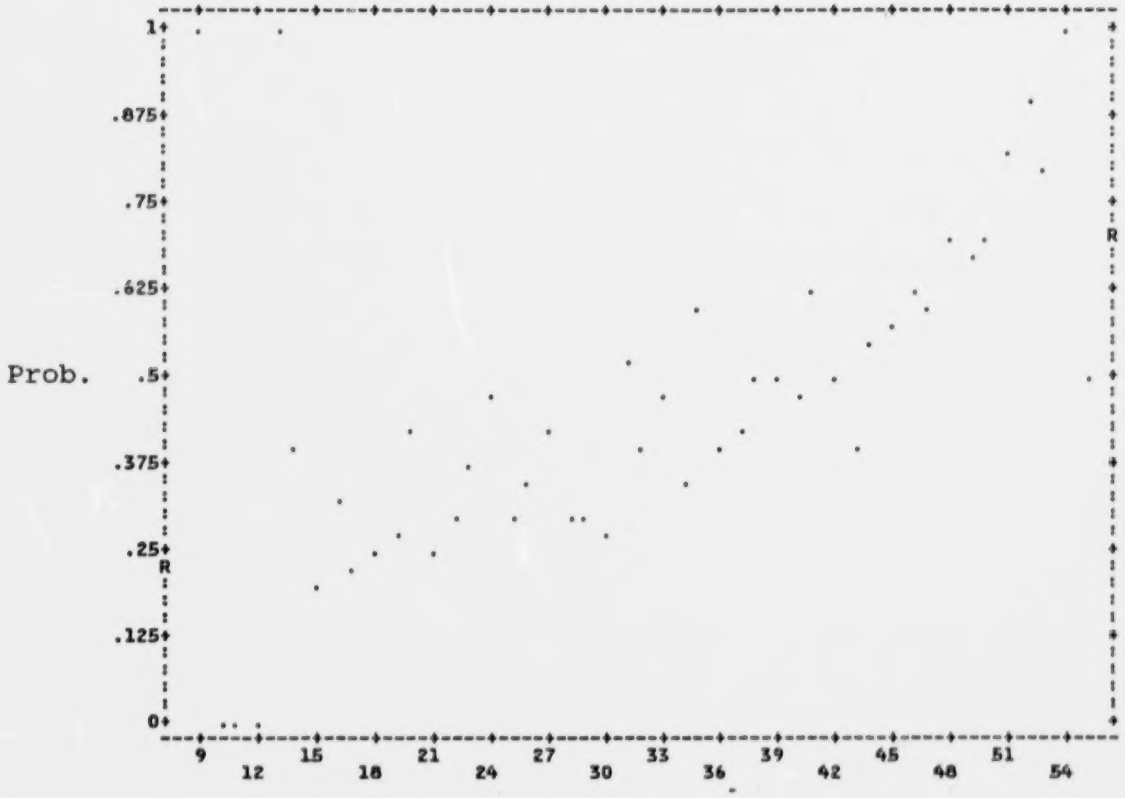

Figure 3. Proportion of correct response (p-value) by total score on a second math subtest indicates item 144 discrimination. Best discrimination is between math subtest scores 42 and 54. The item does not discriminate well at the lower end of the scale. Group B is significantly correlated ($p < .001$) with item score while controlled for achievement level.

Needless to say, this problem potential exists not only in regression bias studies, but also in the computation of standard errors of measurement for cut-score intervals. Ideally, a selection test will maximize discrimination at the cut-score, thus diminishing the likelihood of artifactual bias (both the test and item are likely to discriminate well at the cut score). Unfortunately, not all tests used for making merit groupings will be ideal for the cut-score. Also, the problem may be that the cut-score interval specified is so wide as to render the selected A and B groups unequivalent.

Future Study

This preliminary effort to explore the usefulness of PM reliability, lacked a credible set of real data. Future analysis should include a real data set having cases which match on pre-, post-, and achievement test scores. Results from the simulated data were encouraging with respect to the precision of reliability estimates (size of confidence bounds), but problems with the use of interval means to estimate interval true scores suggests that a regression approach to the estimate of the true scores should be studied. The appearance of negative and widely varying simulated stability coefficients suggests that the data generating process may need to be re-evaluated.

Potential causes of artifactual bias which may afflict PM techniques need to be considered and investigated. Included among these should be the hypothesized effects of discrepant test

and item discrimination where subgroups have differing distributions.

References

Feldt, L.S., and Brennan, R.S. (1989). Reliability. In R.L. Linn (Ed.), Educational measurement, (3rd ed.) (p.123). New York: Macmillan.

Mehrens, W.A., and Lehmann, I.J., (1984). Measurement and evaluation in education and psychology (3rd ed.). (p.272). Chicago: Holt, Rinehart, and Winston.

Thorndike, R.L., (1988). Reliability. In J.P. Keeves (Ed.), Educational research, methodology, and measurement: An international handbook. (p.331). New York: Pergamon

Thorndike, R.L., (1982). Applied psychometrics. (p.237-238). Boston: Houghton Mifflin.